International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-3, Issue-1, March 2014

Approaches for Word Sense Disambiguation – A Survey

Pranjal Protim Borah, Gitimoni Talukdar, Arup Baruah

Abstract: ***Word sense disambiguation is a technique in the field of natural language processing where the main task is to find the correct sense in which a word occurs in a particular context. It is found to be of vital help to applications such as question answering, machine translation, text summarization, text classification, information retrieval etc. This has resulted in excessive interest in approaches based on machine learning which performs classification of word senses automatically. The main motivation behind word sense disambiguation is to allow the users to make ample use of the available technologies because ambiguities present in any language provide great difficulty in the use of information technology as words in human language that occur in a particular context can be interpreted in more than one way depending on the context. In this paper we put forward a survey of supervised, unsupervised and knowledge based approaches and algorithms available in word sense disambiguation (WSD).***

Index Terms: Machine readable dictionary, Machine translation, Natural language processing, Wordnet, Word sense disambiguation.

I.INTRODUCTION

Word sense disambiguation is the task of detecting the meaning of words in a context in a computational paradigm and also differentiating among the senses of words. The solution of a task in WSD is as hard as most of the difficult problems in artificial intelligence and so WSD is often regarded as an AI complete problem [1]. Word sense disambiguation highly depends on knowledge sources like corpora of texts which may be unlabeled or annotated with word senses, machine readable dictionaries, semantic networks etc because the framework of the procedure is that whenever a sentence is given, WSD makes use of more than one knowledge sources to attach the most exact senses with words in the context. The task description of WSD can be formulated as a method of assigning the appropriate sense to all or some words in the text T where T is a sequence of words (w_o, w_1, ..., w_{n-1}) to find the mapping M from words to senses such that $M(k) \subseteq Senses_J(w_k)$ where M(k) is the subset of senses of w_k which are appropriate in the text T and $Senses_J(w_k)$ is the set of senses in dictionary J for word w_k.

Revised Manuscript Received on 30 March 2014.
* Correspondence Author
Pranjal Protim Borah*, Department of Computer Science and Engineering, Assam Don Bosco University, Guwahati, India.
Gitimoni Talukdar, Department of Computer Science and Engineering, Assam Don Bosco University, Guwahati, India.
Arup Baruah, Assistant Professor Department of Computer Science and Engineering, Assam Don Bosco University, Guwahati, India.

The mapping M can assign more than one sense to w_k belonging to T but eventually the most appropriate sense is selected. Thus WSD is a classification task where word senses are the classes and the classification method classifies each occurrence of the word to more than one class based on external knowledge sources and context.
The conceptual model for the word sense disambiguation system is given below in Fig. 1.

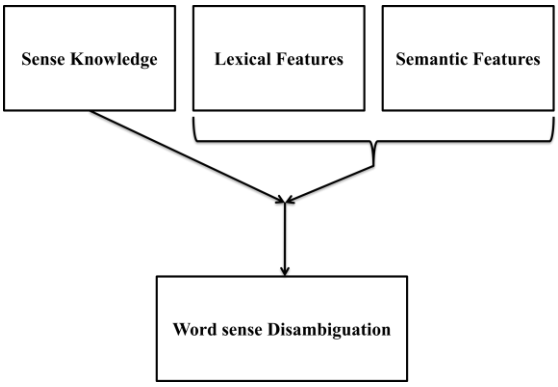

Fig. 1 Conceptual Model for Word Sense Disambiguation

The paper has been further divided into five sections. In section II a brief discussion of knowledge based approaches has been given. In section III supervised disambiguation approach has been highlighted followed by unsupervised disambiguation approach in section IV. In section V elaborations of some evaluation measures for assessing WSD systems is provided and section VI finally concludes our paper.

II.KNOWLEDGE BASED APPROACHES

The idea behind the knowledge based approach is to make extensive use of knowledge sources to decide upon the senses of words in a particular context. It was found that although alternate supervised approaches were more efficient than knowledge based approaches but their advantages also covered a wide range. Collocations, thesauri, dictionaries etc are the most commonly used resources in this approach. Initially knowledge based approaches started in limited domains in 1979 and 1980 [2]. Some of the knowledge based approaches are discussed as follows:

A. Overlap Based Approach

Overlap based approach calls for the requirement of machine readable dictionary (MDR). It includes determination of the different features of the senses of words which are ambiguous along

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

with features of the words in the context. The word sense having the maximum overlap is selected as the appropriate sense in the context. The commonly used algorithms used in overlap based approach are:

1) *WSD using conceptual density: The* conceptual density is the measure of how the concept that the word represents is related to the concept of the words in its context. Conceptual density is related to conceptual distance inversely. The conceptual distance is determined from the wordnet.
2) *Lesk's algorithm:* The Lesk's algorithm used by overlap based approach can be stated as if W is a word creating disambiguation, C be the set of words in the context collection in the surrounding, S be the senses for W, B be the bag of words derived from glosses, synonyms, hyponyms, glosses of hyponyms, example sentences, hypernyms, glosses of hypernyms, meronyms, example sentence of meronyms, example sentence of hypernyms, glosses of meronyms then use the interaction similarity rule to measure the overlap and output the sense which is the most probable having the maximum overlap[3] .
3) *Walker's approach:* Walker's algorithm can be stated as each word is assigned to one or more categories of subjects in the theasurs. Different subjects are assigned to different senses of the word.

B. Selectional Preferences

Selectional Preferences approach imposes restrictions on the possibility of the occurrence of number of meanings of the word in the context. The measure of semantic association is provided by the count of number of instances (W1, W2, Y) in the corpus given by pair of words W1 and W2 occurring in the relation Y. The senses that violate the constraint are omitted. The word sense imposes constraints on the semantic type of the words with which it usually combines grammatically. The semantic appropriateness of word to word can be estimated as the conditional probability of the word W1 given the word W2 [4] as follows-

Y: P(W1/W2,Y) = count(W1,W2,Y)/count(W2,Y).

III.SUPERVISED DISAMBIGUATION

A number of supervised approaches applied to the problem of WSD have reflected the dramatic shift from manually crafted systems to automated machine learning approach. A word expert called the classifier is used to assign appropriate sense to each instance of the concerned word. The classifier learns from the training set containing of a set of examples in which the target word is manually annotated with sense. Some of the common supervised approaches are:

A. Decision Trees

A decision tree divides the training data in a recursive manner and represents the rules for classification in a tree structure. The internal nodes represent test on the features and each branch shows how the decision is being made and the leaf node refers to the outcome or prediction. It is often regarded as a prediction tool. Some popular algorithms for learning decision trees are ID3 and C4.5. On comparison with other machine learning algorithms it was found that several supervised approaches performed better than the decision tree obtained C4.5 algorithm [5]. If the training data is small in size decision tree suffers from prediction unreliability and decision tree also suffers from sparseness of data when there are features with large number of values.

B. Neural Networks

Neural networks processes information based on computational model of connectionist approach. The input includes the input features and the target output. The training dataset is divided into sets which are non-overlapping based on desired responses. When the network encounters new input pairs the weights are adjusted so that the output unit giving the target output has the larger activation. The network can have weights both positive and negative corresponding to correct or wrong sense choice. Neural networks are trained unless the error between the computed and the target output is minimum. Learning in neural networks is eventually updating of weights.

C. Decision Lists

Decision lists contains ordered set of if-then-else rules for assigning category to the test data. Features are obtained from training data which includes rules in the form of (F, S, Score) where F represents feature value, S represents word sense. Rules are arranged in the list as per descending order of the score. For a word w represented in the form of feature vector, the winning sense for the word is the one whose feature has the maximum score in the decision list in matching the input vector.

D. Naive Bayes

Naive Bayes classifier is the classifier based on Bayes theorem and assumption that every feature is class conditionally independent of every other feature. The conditional probability of each sense of the word S_i given the features in the context is calculated to make the final decision.

IV.UNSUPERVISED DISAMBIGUATION

Unsupervised approach unlike supervised approach does not need the beforehand knowledge of sense information in large scale resources for the disambiguation. It is based on the fact that words having similar senses will have similar surrounding words. Word senses are derived by forming clusters of occurrences of words and the task is to classify the new occurrence to the derived clusters. This approach instead of assigning sense labels detects the clusters.

A. Context Clustering

In this approach a vector called context vector is used which is maintained for each word occurrence in the corpora. Clusters are formed for these vectors and each of such clusters corresponds to sense of word to be tested. The initial approach was to have a vector space having words in dimensions. A vector consists of every possible sense of the words. The similarity between two words m and n is provided by the cosine between the respective vectors m and n in a geometrical manner. The grouping of all the vectors gives rise to a co-occurrence matrix.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The matrix may suffer from high dimensionality problem which can be overcomed by the method of latent semantic analysis through singular value decomposition. This is done by merging of dimensions corresponding to words with same senses. Various clustering algorithms are available for sense differentiation as one such mentioned in [6] which was an agglomerative clustering method. In this method at the starting each cluster was having one member corresponding to each instance. The algorithm then continues to group similar pair of clusters as one cluster until a stopping threshold. One more approach was mentioned in [7] called context group discrimination. According to this approach if a word is ambiguous then each occurrence of it is grouped to some sense cluster on the basis of similarity of context of these occurrences. Contextual similarity is measured by the cosine between two vectors corresponding between two words whose similarity is to be determined. Expectation maximization algorithm is used to perform clustering in this approach.

B. Co-occurrence Graphs

In the recent times a certain success is observed in the graph based approaches of unsupervised disambiguation. In a co-occurrence graph the set of vertices V consist of words occurring in text and the set of edges E gives the connection between words co-occurring in the same context. An approach was mentioned in [8] called HyperLex. In this approach nodes are the text words in the paragraph and the edge of the graph signifies that the two words occur in the same paragraph and a weight is given to each edge corresponding to the frequency at which the words connected by the edge co-occur together. The weight can be represented as:

W_{mn}= 1 - Max{ P(W_m/W_n), P(W_n/W_m) } where $P(W_m/W_n)$ represents frequencymn/frequencyn and frequencymn is the frequency at which words W_m and W_n co-occur and frequencyn represents frequency at which W_n occurs within the context. Words having higher co-ocurrence frequency will have weights near about zero and words which co-occur in rare form will have weights near about one.

Another graph based algorithm for deriving word senses is PageRank algorithm which is extensively used in Google search engine. PageRank algorithm can be used to estimate the importance of objects whose relations can be described by a graph.

C. Word Clustering

In this technique words having similar meanings are assigned to the same cluster. One of the approach mentioned in [9] was to find the sequence of words same as the target word. The similarity between the words is given by syntactical dependency. If W consist of words which are similar to w_m then a tree is formed initially with only one node w_m and a node w_i will have a child node w_m when w_i is found to be the word with most similar meaning to w_m. Another approach mentioned in [10] called clustering by committee algorithm represents each word as a feature vector. When target words are encountered a matrix called similarity matrix S_{mn} is constructed whose each element is a similarity between two words w_m and w_n. In the subsequent step of this algorithm committees are formed for a set of words W in recursive manner. The clustering algorithm then tries to find those words not similar to the words of any committee. These words which are not part of any committee are again used to form more committees. In the final step each target word belonging to W will be a member of committee depending on its similarity to the centroid of the committee. The clustering technique used is average-link clustering.

V.PERFORMANCE METRICS

The evaluation measures for assessing a WSD system which is responsible for improving the performance of applications such as machine translation, information retrieval are mentioned below:

Coverage C- Coverage is the measure of percentage of words in the test data for which the WSD system has given sense assignment. It is represented as:

C= Answers provided/Total answers to provide Precision

P- Precision is the measure of ratio of correct answers provided to the answers provided. It is represented as:

P= Correct answers provided/Answers provided

Recall R- Recall is the measure of the ratio of correct answers provided to the total number of answers to provide. It is represented as:

R= Correct answers provided/ Total answers to provide.

F1 measure- F1 measure is the weighted harmonic mean of precision and recall. It is represented as:

F1 measure= (2 * P * R)/(P+R)

Precision is found to be equal to Recall when coverage is 100%.

Table I
COMPARISON OF DIFFERENT SUPERVISED APPROACHES BASED ON ACCURACY

Approach	Average precision	Average baseline accuracy
Naïve Bayes	64.13%	60.9%
Exemplar based	68.6%	63.7%
Decision lists	96%	63.9%
SVM	72.4%	55.2%
Perceptron Trained HMM	67.6%	60.9%

In the above table a comparison of different supervised approaches has been given based on their Average precision and Average baseline accuracy [16].

VI.CONCLUSION

WSD is a very complex task in Natural language processing as it has to deal with complexities found in a language. In this paper we have put forwarded a survey of comparison of different approaches available in word sense disambiguation with primarily focussing on the knowledge based, supervised and unsupervised approaches. We concluded that supervised approach is found to perform better but one of its disadvantage is the requirement of a large corpora without which training is impossible which can be overcame in unsupervised approach as it does not rely on any such large scale resource for the disambiguation.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Knowledge based approach on the other hand makes use of knowledge sources to decide upon the senses of words in a particular context provided machine readable knowledge base is available to apply.

REFERENCES

1. Samit Kumar, Neetu Sharma, Dr. S. Niranjan, "Word Sense Disambiguation Using Association Rules: A Survey", International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 2, Issue 2, 2012.
2. J. Sreedhar, S. Viswanadha Raju, A. Vinaya Babu, Amjan Shaik, P. Pavan Kumar, "Word Sense Disambiguation: An Empirical Survey", International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-2, May,2012.
3. A. Blum and S. Chawla , " Learning from labeled and unla-beled data using graph mincuts", ICML,2001.
4. D. HINDLE and M. ROOTH, "Structural ambiguity and lexical relations", Computat. Ling.19, 1, 103–120, 1993.
5. R. J. MOONEY, "Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning", In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 82–91, 1996.
6. T. PEDERSEN and R. BRUCE, "Distinguishing word senses in untagged text", In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP, Providence, RI), 197–207, 1997.
7. H. SCHUTZE, "Automatic word sense discrimination", Computat.Ling.24, 1, 97–124, 1998.
8. J. V´ERONIS, "Hyperlex: Lexical cartography for information retrieval", Comput. Speech Lang. 18, 3, 223–252, 2004.
9. D. LIN., "Automatic retrieval and clustering of similar words", In Proceedings of the 17th International Conference on Computational linguistics (COLING, Montreal, P.Q., Canada). 768–774,1998.
10. D. LIN and P.PANTEL, "Discovering word senses from text", In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Edmonton, Alta., Canada). 613–619, 2002.
11. A. PURANDAR and T. PEDERSEN, "Improving word sense discrimination with gloss augmented feature vectors", In Proceedings of the Workshop on Lexical Resources for theWeb andWord Sense Disambiguation (Puebla, Mexico), 123–130, 2004.
12. D. YAROWSKY, "Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French", In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (Las Cruces, NM). 88–95, 1994.
13. S. ABNEY and M. LIGHT, "Hiding a semantic class hierarchy in a Markov model", In Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing (College Park, MD), 1–8, 1999.
14. Arindam Chatterjee, Salil Joshii, Pushpak Bhattacharyya, Diptesh Kanojia and Akhlesh Meena,, "A Study of the Sense Annotation Process: Man v/s Machine‖", International Conference on Global Wordnets, Matsue, Japan, January, 2012.
15. M. Nameh, S.M. Fakhrahmad, M. Zolghadri Jahromi, " A New Approach to Word Sense Disambiguation Based on Context Similarity", Proceedings of the World Congress on Engineering 2011 Vol I, July 6 - 8, 2011.
16. Pankaj Kumar, Atul Vishwakarma and Ashwani Kr. Verma, "Approaches for Disambiguation in Hindi Language", International Journal of Advanced Computer Researc, Volume-3 Number-1 Issue-8 March, 2013.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)